Automatic Speech Recognition with Neural Spike Trains

Marcus Holmberg[1], *David Gelbart*[2], *Ulrich Ramacher*[1], *Werner Hemmert*[1]

[1]Infineon Technologies, Corporate Research
Munich, Germany
{firstname.lastname}@infineon.com

[2]International Computer Science Institute
Berkeley, California
gelbart@icsi.berkeley.edu

Abstract

A major difference between the human auditory system and automatic speech recognition (ASR) lies in their representation of sound signals: whereas ASR uses a smoothed low-dimensional temporal and spectral representation of sound signals, our hearing system relies on extremely high-dimensional but temporally sparse spike trains. A strength of the latter representation is in the inherent coding of time, which is exploited by neuronal networks along the auditory pathway. We demonstrate ASR results using features purely derived from simulated spike trains of auditory nerve fibers (ANF) and a layer of octopus neurons. Octopus neurons located in the cochlear nucleus are known for their distinct temporal processing: they not only reject steady-state excitation and fire on signal onsets but also enhance the amplitude modulations of voiced speech. With multi-condition training we do not reach the performance of conventional mel-frequency cepstral coefficients (MFCC) features. With clean training however, our spike-based features performed similarly to MFCCs. Further, recognition scores in noise were improved when features derived from ANFs, which mainly represent spectral characteristics of speech signals, were combined with features derived from spike trains of octopus neurons. This result is promising given the relatively small number of neurons we used and the limitations in how the auditory model was interfaced to the ASR back end.

1. Introduction

Neurons in the auditory system code and process both temporal and spectral information in sound. Spectral information is carried mainly by the “rate–place code”. This code is a result of the spectral decomposition along the length of the inner ear, where a given neuron carries information about a specific frequency range. The stronger the signal is in that frequency range, the higher the firing rate. The rate-place code of the auditory nerve fibers (ANFs) has many similarities with standard mel-spaced spectrogram representations of speech. In addition to the rate, nerve fibers code temporal information by eliciting spikes with exact timing relative to sound signals. This code carries temporal cues such as amplitude modulation and precise onset times which are likely to be important for speech recognition. A weakness of ASR technology today is that temporal information is not exploited as it is in our auditory system, although first attempts have demonstrated that temporal information provides superior recognition results in noisy conditions [10, 9, 11, 6].

In this paper we focus on cochlear nucleus octopus neurons which are known for their distinct temporal processing capabilities [1]. The cochlear nucleus is the first neuronal processing stage and receives inputs directly from auditory nerve fibers (ANFs). Unlike most other neurons, which exhibit sustained activity to continuous excitation, octopus neurons show onset inhibitory responses: they only fire on signal onsets. As they receive predominantly excitatory inputs, their membrane properties are thought to be responsible for suppressing sustained responses [1, 8]. Octopus neurons exhibit an extraordinarily fast membrane time constant (0.3 ms); they detect temporal coincidence of incoming spikes (firing only on synchronous activity of multiple ANFs innervating them) and thus greatly enhance the precision of timing relative to a single ANF [3]. Their electrical behavior is dominated by a low-threshold potassium channel (K_{LT}) with activation kinetics in the order of 2 ms, which is already activated at rest [1]. When their membrane is depolarized, they elicit an initial action potential, but thereafter K_{LT} compensates input currents and keeps the membrane potential below spiking threshold.

2. Auditory model

In this paper we combine our realistic inner ear model [2, 4], which codes sound signals into spike trains of the auditory nerve, with a layer of modeled octopus neurons.

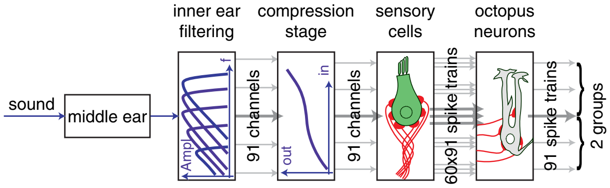

Figure 1: Schematic figure of the model. Sound signals are separated into 91 frequency channels. From one sensory cell 60 spike trains of ANFs drive a single octopus neuron.

2.1. Modeling sound coding into nerve action potentials

The model of the peripheral hearing system consists of a simplified middle ear, a model of inner ear hydrodynamics followed by a compression stage, and a sensory cell model. Here we describe the model very briefly, more details can be found in [4]. BM vibrations were calculated with a computationally efficient wave-digital filter model comprising of 100 sections at a sampling rate of 48 kHz. Our inner ear model covers the complete human hearing range up to 20 kHz; as the sound files used in this paper are bandlimited to 8 kHz, we discarded 9 high-frequency channels and processed only the remaining 91 channels. In this paper we corrected for group delays introduced by the inner ear model to avoid feeding frequency-specific delays to the speech recognition back-end. The sensory cells and synaptic mechanisms were modeled according to [12]. Each sensory cell was connected to multiple auditory nerve fibers (ANFs), which elicited all-or-nothing nerve action potentials. We employed a simplified spike generation module, which still included both absolute and relative refraction. In this paper we only used spike trains from high-spontaneous rate (HSR) ANFs.

Our model provides large dynamic compression of more than 60 dB and generates realistic ANF responses which code sound signals with great fidelity. It replicated bandwidths of human threshold tuning curves and latest measurements of dynamic range compression with great precision [4].

2.2. Temporal processing of octopus neurons

We used a single-compartment model with Hodgkin-Huxley-type ion channels of the octopus neurons [3]. Rothman and Manis measured the properties of the main conductances [8] and derived both steady-state and dynamic equations. We solved the differential equations of the ionic channels by replacing them with difference equations and iterating in the time domain.

We connected 60 auditory nerve fibers from a single frequency channel of our inner ear model to an octopus neuron. We used only excitatory synapses. An action potential was elicited when at least (depending on previous activity) 25% of the input fibers fired synchronously; a spike of an octopus neuron was counted only when its membrane potential crossed 0 mV.

2.3. Spiking representations of sound

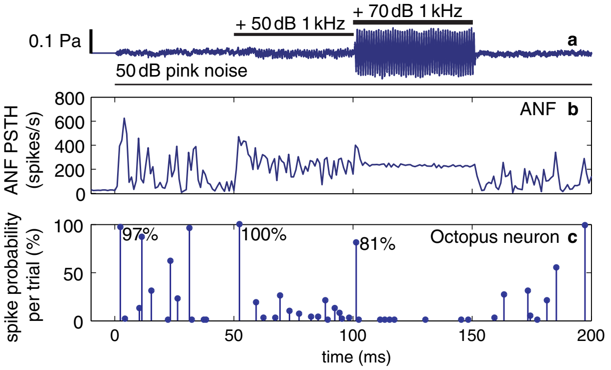

Figure 2: Onset processing of octopus neurons. (a) Sound stimulus is pink noise with a 1 kHz pure tone with stepwise increasing amplitude (rise times: 1 ms). (b) Poststimulus-time histogram (PSTH) of a single ANF. (c) Octopus neuron activity per trial. Spike counts are collected in 1 ms time bins. Spike counts are from 60 ANFs innervating a single octopus neuron summed over 100 trials.

Figure 2 shows the response histogram of an ANF and an octopus neuron with characteristic frequencies of 1 kHz. Sound stimulus was pink noise (level: 50 dB SPL, switched on at t=0 s) to which a 1 kHz pure tone was added (50 dB SPL for 50 ms<t<100 ms; 70 dB SPL for 100 ms<t<150 ms; all levels are root-mean square values, no filtering applied). Before a signal is applied (t<0 s), the ANF fires with its spontaneous rate of approximately 30 spikes/s. When the noise signal is switched on, the ANF reacts with an initial sharp rise of its firing rate which then exhibits strong fluctuations following the temporal fine-structure of the noise. When the pure tone is switched on, another transient is generated. During the tone the sustained rate is approximately 226 spikes/s (averaged over 70 ms<t<100 ms). After the tone's level is raised, the sustained firing rate hardly increases any more (232 spikes/s; 120 ms<t<150 ms); HSR fibers saturate about 40 dB above threshold. Still, the fiber is able to generate a transient onset response. Note that the fluctuations of the ANF activity decreases as the level of the pure tone increases.

Figure 2c shows the reaction of an octopus neuron to the same signal. At the onset of the pink noise, the neuron fired with 97% probability (100 trials); when the pure tone was added, it always fired. During the duration of the tone burst, octopus neurons suppressed further spikes. When the level of the tone was increased from 50 to 70 dB, the neuron fired with a probability of 81%. Note that octopus neuron activity was also depressed during the duration of the tone: during the pink noise signal the octopus neuron fired approximately 3 times per trial (excluding reaction to the onset); this rate decreased to 1.4 spikes/trial (50 dB tone) and 0.1 spikes/trial (70 dB tone). After the tone is switched off, octopus neuron activity slowly recovers.

Figure 3 shows neuronal responses to the first part of the spoken utterance "*a*" (ISOLET msa0-A1-t.sph) without background noise (left column) and with pink noise added with a signal-to-noise ratio of 15 dB (right column). The excitation pattern of HSR auditory fibers along the cochlea is shown in panels b+g. The responses of sixty fibers per frequency channel were modeled and plotted. Darker areas code higher firing rates. The spoken utterance starts at approximately t=95 ms. The inner ear performs a frequency decomposition, and thus we can read the excitation pattern as a spectrogram. Three formant regions can be discerned (F1 around 600 Hz, F2-F3 around 2 kHz, and F4 at 4 kHz), as well as the resolved fundamental frequency around 180 Hz. (Compare also Fig. 3, panels d+i which show the ANF responses with a reduced temporal resolution. See Sec. 3 for further details.) The formant transitions typical for the utterance "*a*" (F1 moving downwards in frequency with time and F2 slightly upwards) starts at around 200 ms. The whole course of the transition is not shown.

In addition to the spectral information, the ANFs visibly carry temporal information. Each stroke of the glottis (each 5.6 ms in the first 50 ms of the speech signal, corresponding to a fundamental frequency of 178 Hz) generates synchronous action potentials in all formant regions.

The highly synchronous ANF firing in response to speech across frequency channels stands in contrast with the irregular spontaneous firing seen for the quiet period (panel b, 50–90 ms) and the response to pink noise stimuli only (panel g, 50–90 ms).

The octopus neurons (panels c+h) further highlight this difference. Note that Fig 3b+g show the response of 60 ANFs per frequency channel, whereas there is only *one* octopus neuron per frequency channel. Nevertheless, octopus neurons code each glottis stroke in the region between 0.6 and 1 kHz (first formant) and the resolved fundamental (180 Hz) with high fidelity in both clean and noisy conditions. Thus they not only respond to signal onsets, but also to amplitude modulated stimuli.

We found further insight into the temporal information processing properties of an octopus neuron by applying the spike-triggered reverse-correlation technique [3]. In summary, temporal processing of an octopus neuron very much resembled a band-pass filter with a pass-band (-3 dB) reaching from 110 Hz to 650 Hz. Notably, the high sensitivity of octopus neurons to amplitude modulations above 100 Hz coincides with the fundamental frequency of speech sounds. Octopus neurons fire preferentially at onsets of sound signals which provides a good means of sound segmentation. Octopus neurons discern voiced

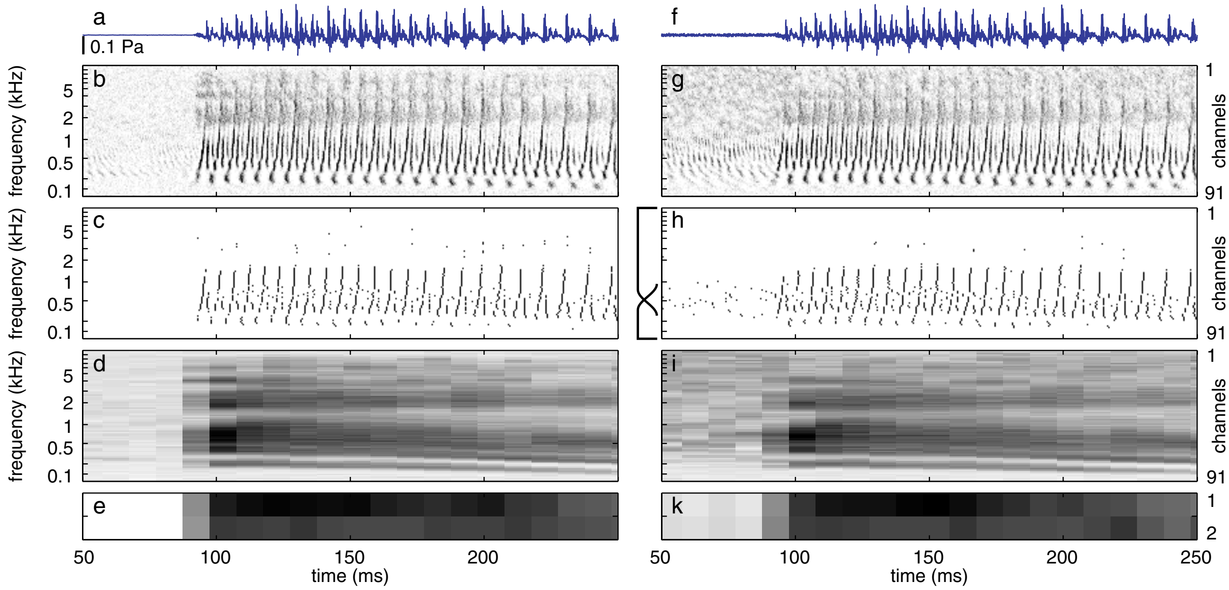

Figure 3: Neuronal firing patters to the spoken utterance "*a*" with no background noise (panel a) and with pink noise (15 dB SNR, panel f). (b+g) Auditory nerve excitation pattern. (c+h) Octopus neuron responses. (d+i) ANF responses with reduced temporal resolution, as used for ASR. (e+k) Features derived from octopus neuron responses as used for ASR. Spectral integration was achieved by using the windows indicated to the left of panel (h). Responses shown are for 60 ANFs and one octopus neuron (innervated by the 60 ANFs) per channel. Spikes were collected in 0.5 ms time bins. Note that in this paper we corrected for group delays introduced by the inner ear model.

sounds, to which they generate pitch-synchronized action potentials, and noise-like signals, which only elicit sporadic, non-synchronized activity.

3. Automatic speech recognition task

We used a noisy speech corpus based on OGI's ISOLET corpus[1]. ISOLET is about 1.25 hours of audio recorded in a quiet environment; each utterance is the name of a letter of the English alphabet. We artificially added one of eight different noises from the RSG-10 collection [13] to each utterance at six different signal-to-noise ratios (including a "clean" case in which no noise was added). ISOLET comes divided into five equal-sized parts (ISOLET1 through ISOLET5). We always used four of the five parts as training and the remaining part as test to achieve five-way cross-validation. Three of the noise types were added to all five parts of ISOLET, the remaining five noises were used for one part each. This means that for each of the five different divisions of ISOLET into training and test data that we use, there are three matched noises (found in both training and test) and one mismatched noise (found only in test).

We refer to this as *multicondition training*. We also used a *clean training* case in which no noise was added to the training utterances. The experimental results we report are averaged over the five different divisions of the corpus utterances, but separated into the clean- and a multi-condition training case.

Our recognizer back end was built with Cambridge's HTK, and used a different whole-word HMM for each letter. It is derived from the back end configuration used in [5]. For greater training robustness, we reduced the number of model parameters compared to [5], using four-state HMMs with each state modeled by a mixture of three diagonal-covariance Gaussians. We did not use endpointing. Using Mel-frequency cepstral coefficient (MFCC) features, this system had a 92.7% accuracy on the original ISOLET corpus.

3.1. Interfacing to the speech recognizer

Conventional Hidden-Markov-Model (HMM) ASR back ends are not suited to process spike trains, specifically their high temporal resolution. We therefore had to reduce the dimensionality of our data. We temporally integrated the spike trains (of both ANFs and octopus neurons) for each model channel using a 25 ms Hanning window which we advanced in steps of 10 ms. Fig. 3, panels d+i show the resulting ANF features. To reduce the number of ANF features in the frequency domain we applied a DCT and kept the first 12 or 14 cepstral coefficients (including C0). This dimensionality reduction scheme removes all fine temporal resolution inherent in spike trains. Our features are therefore extracting the rate-code of ANFs only. The outputs of the 91 octopus neurons were spectrally integrated to produce two features (Fig. 3e+k) using the window functions shown just left of Fig. 3h. In order to better fit the assumptions of the Gaussian mixtures, we post-processed our auditory model features using a histogram-based Gaussianization transform and (usually) a Karhunen-Loeve Transform (KLT) [2]. We then augmented our feature vector by adding first- and second-order temporal derivatives (calculated over nine frames centered on the current frame). For comparison, we calculated MFCCs, using 12 or 14 cepstral coefficients. MFCCs were augmented with first- and second-order temporal derivatives in the same

[1] For more information about our noisy ISOLET corpus, see http://www.icsi.berkeley.edu/Speech/papers/eurospeech05-onset/isolet. We are working on making a package of scripts and configuration files available at that location so that others can use our noisy ISOLET corpus and our recognizer back end.

[2] We applied Jeff Bilmes' pfile_gaussian and pfile_klt tools available at: http://www.icsi.berkeley.edu/~dpwe/projects/sprach/sprachcore.html.

manner as for the auditory features. We did not Gaussianize MFCCs.

3.2. Recognition results

Table 1: *Summary of recognition results. Features derived from spike trains of the auditory nerve (ANF) and octopus neurons (O) are compared with MFCCs.*

feature type (dimension)	dimension after KLT	clean training	multicondition training
ANF (12)	No KLT	48.6	70.3
ANF (12)	12	42.4	69.7
ANF (14)	12	51.7	67.7
ANF (12) + O (2)	12	55.2	69.7
MFCC (12)	No KLT	47.8	77.7
MFCC (12)	12	56.4	74.6
MFCC (14)	12	56.9	74.1

Table 1 compares recognition accuracies using various features derived from spike trains to accuracies for MFCC features. We first discuss our results with clean training. Without KLT the recognition accuracy for the 12 features derived from ANF spike trains (48.6%) was comparable to that for the MFCCs (47.8%). With KLT, the MFCCs improved to 56.4% accuracy in clean training, while the 12 ANF features worsened to 42.4%.

Adding two features derived from octopus neurons to the 12 ANF features–keeping the feature vector length prior to adding temporal derivatives constant at 12, by using the KLT for dimensionality reduction–improved the accuracy with auditory features to 55.2%. As a control experiment we tried 14 ANF features (reduced to 12 by KLT), which improved performance compared to 12 ANF features alone but was still outperformed by the combined ANF/onset features. Morris and Pardo [7] discuss phonetic information carried by onsets.

With multicondition training, the use of KLT never improved recognition accuracy, and accuracy with features derived from spike trains did not reach the performance of MFCCs.

4. Conclusion

In this paper we demonstrate speech recognition using realistic spike trains from ANFs and from octopus neurons. Much of the temporal information in ANF spike trains was destroyed by the dimensionality reduction we employed in interfacing to the HMM back end; the resulting features represent rate coding of the sound spectrum. Our recognition accuracy results for clean training show that temporal information extracted by octopus neurons can complement this predominantly spectral information.

Our results are promising given the fact that only a relatively small number of ANF spike trains were used (5,500 nerve fibers compared to about 100,000 estimated for humans) and that we only included HSR fibers in this study. HSR fibers predominantly code sound close to hearing threshold and approximately 40 dB above; at higher levels they saturate. Extending the modeling to low spontaneous rate fibers might improve recognition accuracy. We also are interested in having the simulated octopus neurons take input from more than one frequency channel of our inner ear model, so that they can respond to common onsets in different frequency channels.

The human auditory system deals with a huge number of "features" in multiple parallel and hierarchical layers. This highlights a major dilemma in ASR research: it is an open problem how best to handle such a high-dimensional space or take advantage of the excellent temporal precision inherent in neuronal spike trains. We feel that attention-based mechanisms which focus on relevant features and discard others might prove a useful tactic in bringing ASR closer to human performance.

4.1. Acknowledgements

We want to thank J. Bilmes, O. Cetin, P. Hix, H. Hermansky, H.-G. Hirsch, M. Karnjanadecha, N. Morgan, H. Steeneken, C. Wooters and S. Zahorian for their help. This work was funded by the German Federal Ministry of Education and Research (reference number 01GQ0443).

5. References

[1] Ferragamo, M. J. and Oertel, D., "Octopus cells of the mammalian ventral cochlear nucleus sense the rate of depolarization", J. Neurophysiol. 87 (2002) 2262–2270.

[2] Hemmert, W., Holmberg, M. and Gelbart, D., "Auditory-based automatic speech recognition", in: Proc. of SAPA Workshop, ICSLP-Interspeech (2004) SII, pp. 1–6.

[3] Hemmert, W, Holmberg, M. and Ramacher, U., "Temporal sound processing by cochlear nucleus octopus neurons", In: Proceedings of the International Conference on Artificial Neuronal Networks (ICANN), 2005, in press.

[4] Holmberg, M. and Hemmert, W., "An auditory model for coding speech into nerve-action potentials", in: Proceedings of the Joint Congress CFA/DAGA (2004) 773–4.

[5] Karnjanadecha, M. and Zahorian, S.A., "Signal modeling for high-performance robust isolated word recognition", IEEE Trans. Speech and Audio Proc., 9(6):647–654, 2001

[6] Kim, D.-S., Lee, S.-Y. and Kil, R.M., "Auditory processing of speech signals for robust speech recognition in real-world noisy environments", IEEE Trans. Speech and Audio Proc. 7(1): 55-69, 1999.

[7] Morris, A.C. and Pardo, J.M., "Phoneme transition detection and broad classification using a simple model based on the function of onset detector cells found in the cochlear nucleus", Eurospeech, 115–118, 1995.

[8] Rothman, J. S. and Manis, P. B., "The roles potassium currents play in regulating the electrical activity of ventral cochlear nucleus neurons", J. Neurophysiol. 89 (2003) 3097–3113.

[9] Sandhu, S., Ghitza, O. and Lee, C.-H., "A comparative study of mel cepstra and EIH for phone classification under adverse conditions", IEEE ICASSP, 409–412, 1995.

[10] Seneff, S., "A joint synchrony/mean-rate model of auditory speech processing", J. Phonetics, 16: 55–76, 1988.

[11] Sheikhzadeh, H. and Deng, L., "Speech analysis and recognition using interval statistics generated from a composite auditory model", IEEE Trans. Speech and Audio Proc., 6(1): 90–94, 1998.

[12] Sumner, C. J., Lopez-Poveda, E. A., O'Mard, L. P. and Meddis, R., "A revised model of the inner-hair cell and auditory-nerve complex", J. Acoust. Soc. Amer. 111 (2002) 2178–2188.

[13] Steeneken, H.J.M. and Geurtsen, F.W.M., "Description of the RSG-10 noise data-base", Report IZF 1988–3, TNO Institute for Perception, The Netherlands, 1988.